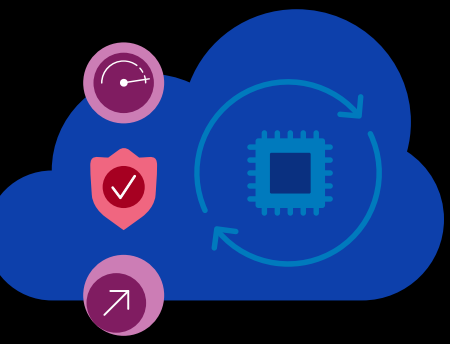
DATA SHEET



F5 BIG-IP Next for Kubernetes Integration with NVIDIA RTX PRO™ Server

Key benefits

Enhance efficiency
Leverage capabilities such as intelligent routing and load balancing to deliver AI performance with optimized costs, maximized GPU utilization, and minimized token latency.

Strengthen security
Deploy zero trust capabilities at the edge and safeguard multi-tenancy to enable secure AI at scale.

Unlock scalability
Enable scalable AI innovation, even in power-constrained data centers, by offloading AI workloads to NVIDIA BlueField-3 DPUs and RTX PRO GPUs.

Maintain control
Take advantage of granular control to manage per-tenant bandwidth, prevent resource contention, and ensure consistency.

Improve experiences
Reduce latency and strengthen reliability to deliver seamless AI experiences and enhanced results.

Enterprise-grade, secure, optimized north-south traffic management

The integration of F5® BIG-IP Next for Kubernetes with NVIDIA RTX PRO™ 6000 Blackwell Server Edition expands the collaboration between F5 and NVIDIA to enhance high-performance traffic management, intelligent security, and scalability for advanced AI applications. The integration combines advanced security and application delivery technologies from F5 with industry-leading NVIDIA GPU innovation, enabling organizations to accelerate forward progress while securing distributed AI workloads.

The NVIDIA RTX PRO 6000 Blackwell Server Edition is designed to address the most demanding enterprise and industrial AI workloads with embedded data processing acceleration and exceptional computational capabilities for multimodal AI inference, scientific computing, graphics, and more. Incorporating the NVIDIA BlueField®-3 DPU, the solution further elevates advanced networking, security enforcement, and governance at the infrastructure level, providing greater operational control for enterprise AI environments.

F5 has optimized its BIG-IP Next for Kubernetes platform to seamlessly integrate with NVIDIA's BlueField-3 DPUs, enabling enhanced performance for AI applications. This integration enables more efficient token generation, smart routing for large language models (LLMs), and improved load balancing of inferencing endpoints. By leveraging the networking capabilities of NVIDIA BlueField-3 DPU, this solution delivers the scalability, performance, security, and workload isolation critical for enterprise AI deployments.

Further strengthening the capabilities of this collaboration is the NVIDIA Enterprise Reference Architecture (Enterprise RA), which is optimized for multi-node AI and hybrid applications. The RA utilizes a 2-8-5-200 node architecture, deploying RTX PRO 6000 Blackwell Server Edition GPUs. It is powered by the NVIDIA Spectrum-X Networking Platform, offering advanced networking features tailored specifically for AI workloads.

The Enterprise RA is a modular architecture built on NVIDIA-Certified™ Systems, each configured with eight RTX PRO 6000 Blackwell Server Edition GPUs for optimal performance. Scaling in four-node scalable units (SU), the architecture can grow to encompass up to 32 NVIDIA-Certified Systems, supporting a total of 256 RTX PRO 6000 Blackwell Server Edition GPUs—delivering unmatched scalability and flexibility for enterprise-grade AI implementations.

Key features

High-efficiency token generation

- Maximizes token generation per watt for resource optimization
- Supports high-throughput processing for LLMs
- Enhances AI inference efficiency with reduced latency

Intelligent application traffic management

- Provides dynamic routing for AI-accelerated applications
- Delivers optimized load balancing for multimodal workloads
- Ensures seamless scaling across distributed environments

Comprehensive L3-L7 security capabilities

- Provides advanced security features including L3 firewall, L4-L7 firewall, and DDoS protection
- Enforces encryption, authentication, and rate-limiting policies directly on the DPU for high-performance security
- Secures sensitive data and AI workloads across Kubernetes clusters
- Protects AI environments from evolving cybersecurity threats using robust access control and traffic management

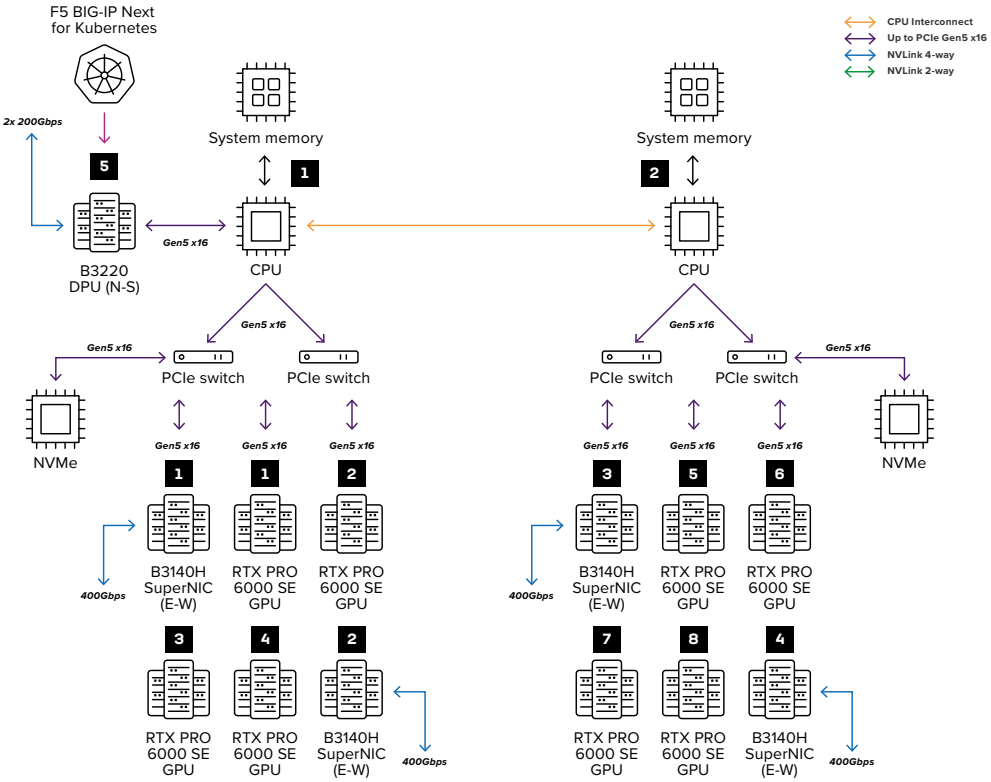


Figure 1: 8 GPU System Configuration (2-8-5-200)

Deploy high-performance traffic management

The integration of F5 BIG-IP Next for Kubernetes with NVIDIA RTX PRO Server ensures seamless traffic management, reducing latency and offloading security, ingress and egress load balancing to NVIDIA BlueField-3 DPU for higher efficiency. Enterprises can efficiently route LLMs and multimodal AI inference tasks, optimizing performance and scalability for demanding AI applications. By streamlining complex workloads, organizations can deliver faster innovation across distributed environments without compromising performance.

Implement zero trust security at the edge

F5 BIG-IP Next for Kubernetes leverages enterprise-grade security capabilities to safeguard sensitive AI workloads, ensuring high-performance and secure operations in modern AI environments. Running directly on the DPU, the solution enforces encryption, authentication, rate-limiting policies, and mutual TLS, all without consuming host CPU or GPU cycles. Per-tenant access control and DOCA-offloaded DDoS protection further enhance application security and scalability. These features protect AI-powered applications at scale and help organizations build trust in their deployments.

Key features continued

Programmable multi-cluster partition security

- Prevents data leakage across multiple Kubernetes clusters
- Enables secure partitioning of distributed AI workloads
- Improves compliance management for enterprises

Optimized performance for NVIDIA RTX PRO Server

- Enhances application delivery for NVIDIA RTX PRO Server workloads
- Optimizes LLM and AI inference routing for GPU efficiency
- Reduces latency for multimodal AI tasks and scientific computing
- Maximizes throughput for distributed AI applications at scale

Seamless integration with Kubernetes environments

- Features F5 Lifecycle Operator (FLO)—a Kubernetes-native OLM operator—streamlining installation, upgrades, and policy synchronization
- Simplifies lifecycle management with declarative, GitOps-based workflows
- Supports advanced traffic management with Kubernetes Gateway API integration for containerized workloads
- Enables easy deployment within existing cloud-native ecosystems, reducing complexity for enterprises adopting AI infrastructure

Implement zero trust security at the edge

F5 BIG-IP Next for Kubernetes leverages enterprise-grade security capabilities to safeguard sensitive AI workloads, ensuring high-performance and secure operations in modern AI environments. Running directly on the DPU, the solution enforces encryption, authentication, rate-limiting policies, and mutual TLS, all without consuming host CPU or GPU cycles. Per-tenant access control and DOCA-offloaded DDoS protection further enhance application security and scalability. These features protect AI-powered applications at scale and help organizations build trust in their deployments.

Maximize efficiency for energy-constrained data centers

The combined solution optimizes AI infrastructure performance and efficiency by offloading Layer 4 to Layer 7 ingress and egress load balancing, token governance, and intelligent LLM routing to NVIDIA BlueField-3 DPUs and RTX PRO GPUs. By streamlining these networking tasks, enterprises can achieve faster and more efficient AI operations while reducing the computational burden on host systems. This energy-efficient approach minimizes operational costs and enables scalable AI acceleration for power-constrained data centers. Organizations can future-proof their infrastructure to support next-generation AI processing demands without compromising on efficiency or performance.

Improve GPU utilization and token efficiency optimization

BIG-IP Next for Kubernetes provides AI-optimized intelligent load balancing to maximize GPU utilization and minimize token latency. By leveraging metrics such as token throughput, time to first token, and cost per token, BIG-IP Next for Kubernetes dynamically adapts traffic routing in real time to sustain model efficiency and reduce inference costs.

Protect multi-tenancy and enhance service governance

BIG-IP Next for Kubernetes enforces robust logical isolation to safeguard multi-tenant environments, ensuring secure and efficient operations across tenants, namespaces, and AI services. It enables fine-grained control with per-tenant bandwidth guarantees, preventing resource contention and maintaining performance consistency. The solution also supports certificate segmentation, ensuring cryptographic assets are isolated and uniquely managed for each tenant.

More information

To learn more about F5 BIG-IP Next for Kubernetes, visit f5.com.

Blog: F5 unleashes innovation with powerful new AI capabilities on BIG-IP Next for Kubernetes on NVIDIA BlueField-3 DPUs

Solution overview: Powering GPUaaS and AI Inferencing services with F5 and NVIDIA

Solution overview: Driving AI business outcomes with intelligence and security at scale

DC 10.2025 | JOB-CODE-1771678146